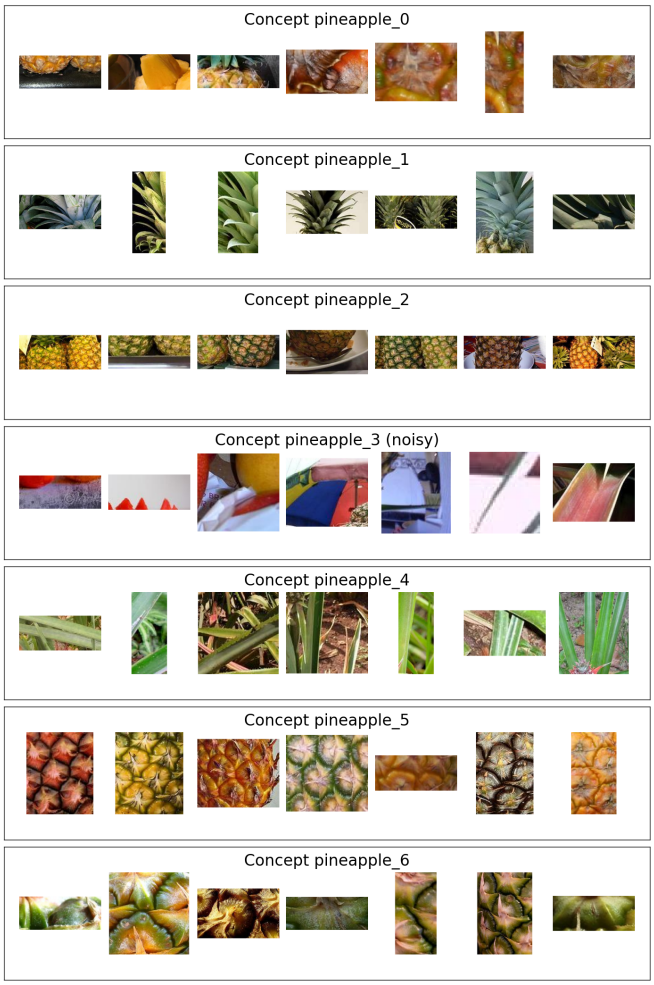
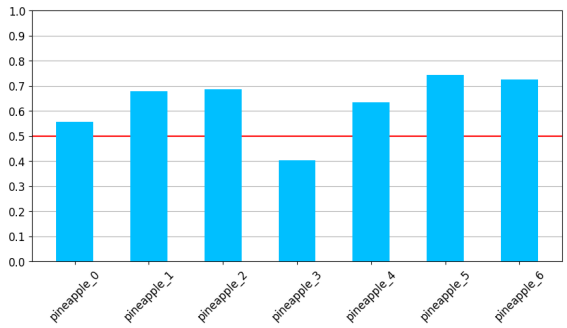
Activation-Based Concept Extraction for Explainability in Image Classification

Supplementary Material

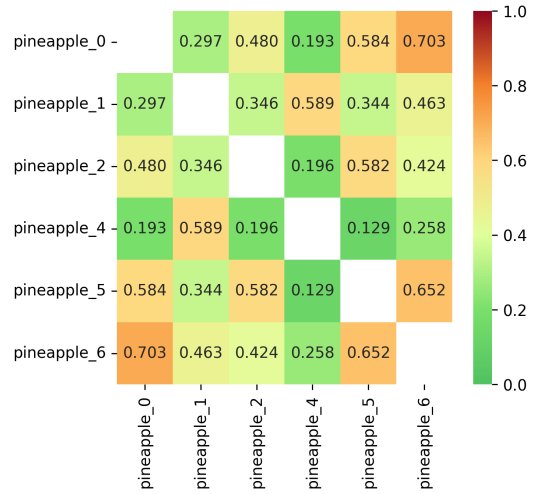
6. Examples of Concept Filtering and Merging



(a) Concepts

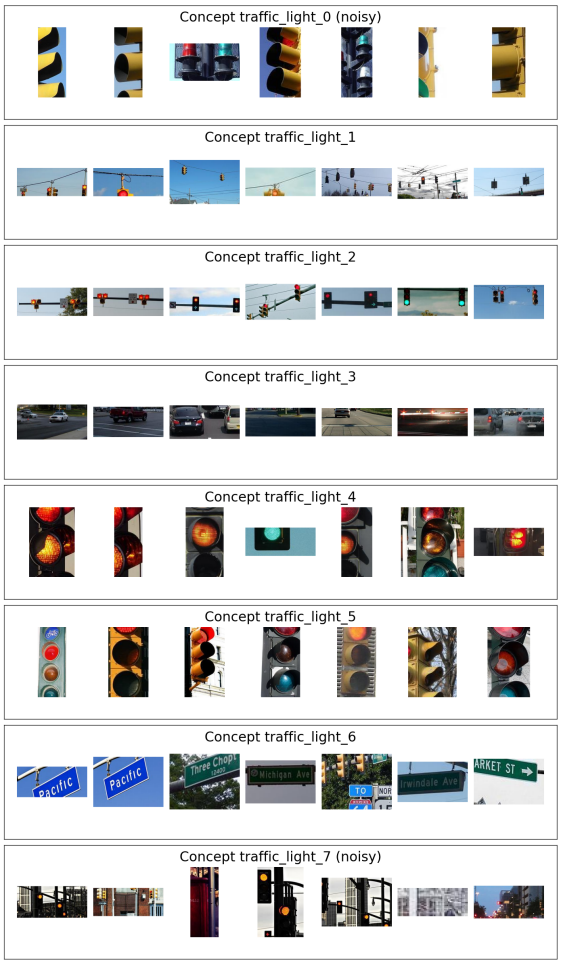


(b) Internal similarity scores for all extracted concepts.

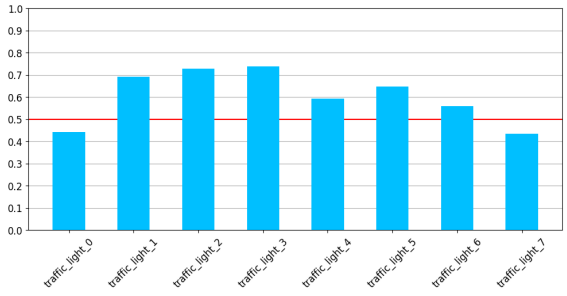


(c) Similarity matrix between CAV directions for non-noise concepts.

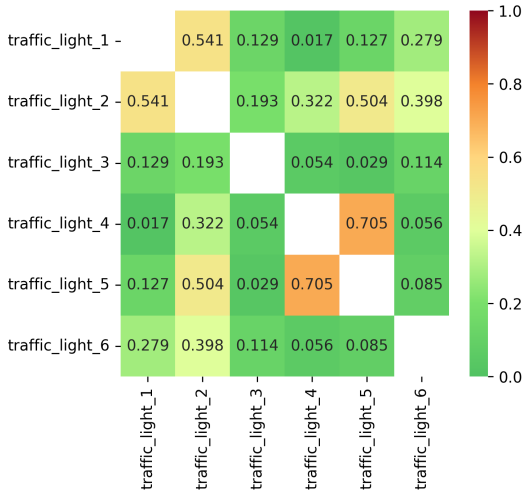
Figure 8. Concepts extracted for class “pineapple” with ResNet50, along with the internal similarity score for each extracted concept and the similarity matrix showing similarity scores between each pair of distinct concepts. Concept “pineapple_3” is the only one falling below the noise threshold, which is coherent with our perception as it is hard to find a common concept between its images. We can also see that the similarity matrix aligns with our perception, for example marking concepts “pineapple_0”, “pineapple_5” and “pineapple_6” as similar. On the other hand concepts “pineapple_1” and “pineapple_4”, which contain images of different types of leaves, are distinct from concepts representing pineapple peel but similar to each other.



(a) Concepts



(b) Internal similarity scores for all extracted concepts.



(c) Similarity matrix between CAV directions for non-noise concepts.

Figure 9. Concepts extracted for class "traffic light" with VGG-16, along with the internal similarity score for each extracted concept and the similarity matrix showing similarity scores between each pair of distinct concepts. We can see that concepts "traffic_light_0" and "traffic_light_6" are not very cohesive, which is reflected in their internal similarity scores as they are the noisiest ones. Additionally, concepts "traffic_light_4" and "traffic_light_5" are marked as the most similar pair, which aligns with our perception as they both represent traffic light bodies, although at different levels of zoom.

7. Examples of Concept Extraction

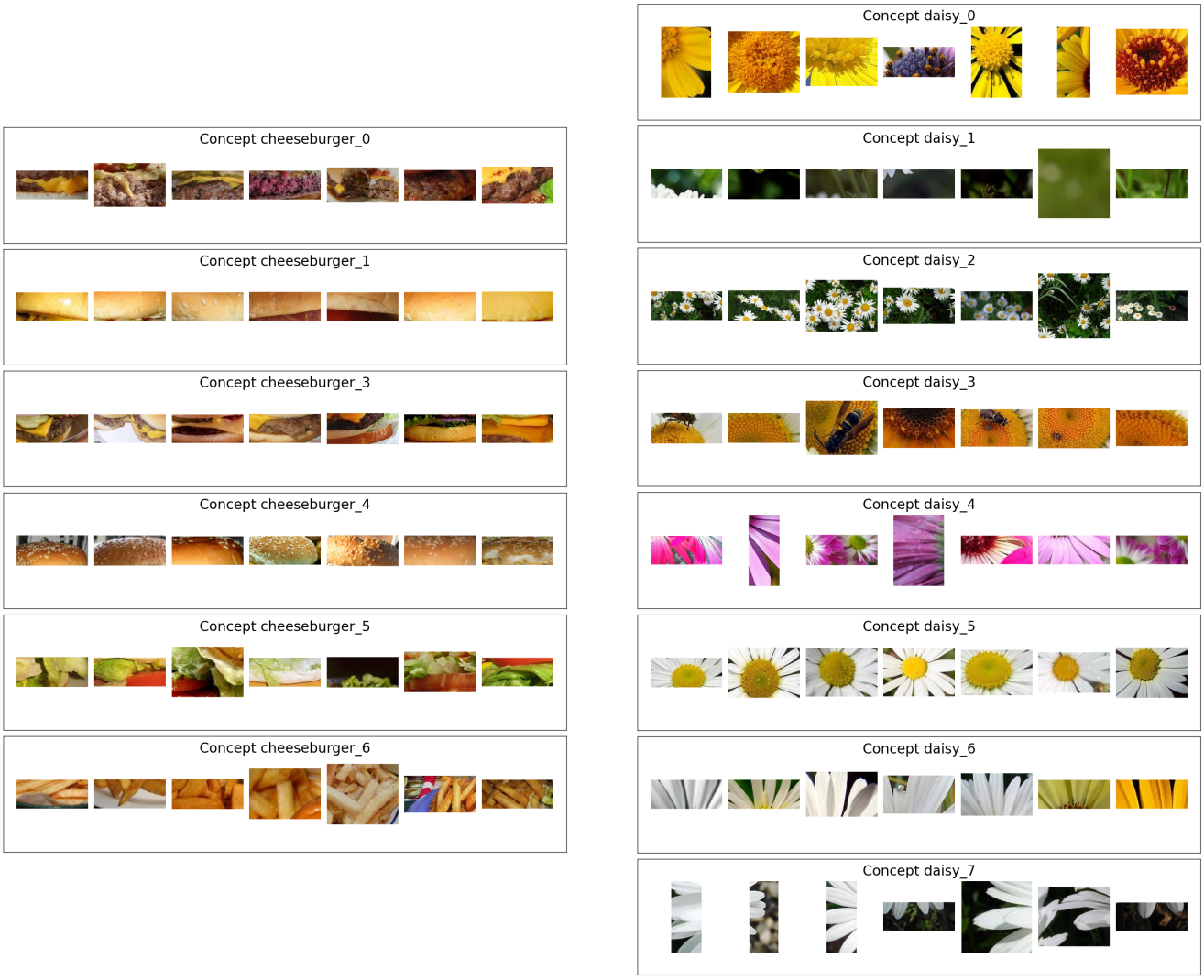


Figure 10. Concepts extracted for class “cheeseburger” and “daisy” with ResNet50.

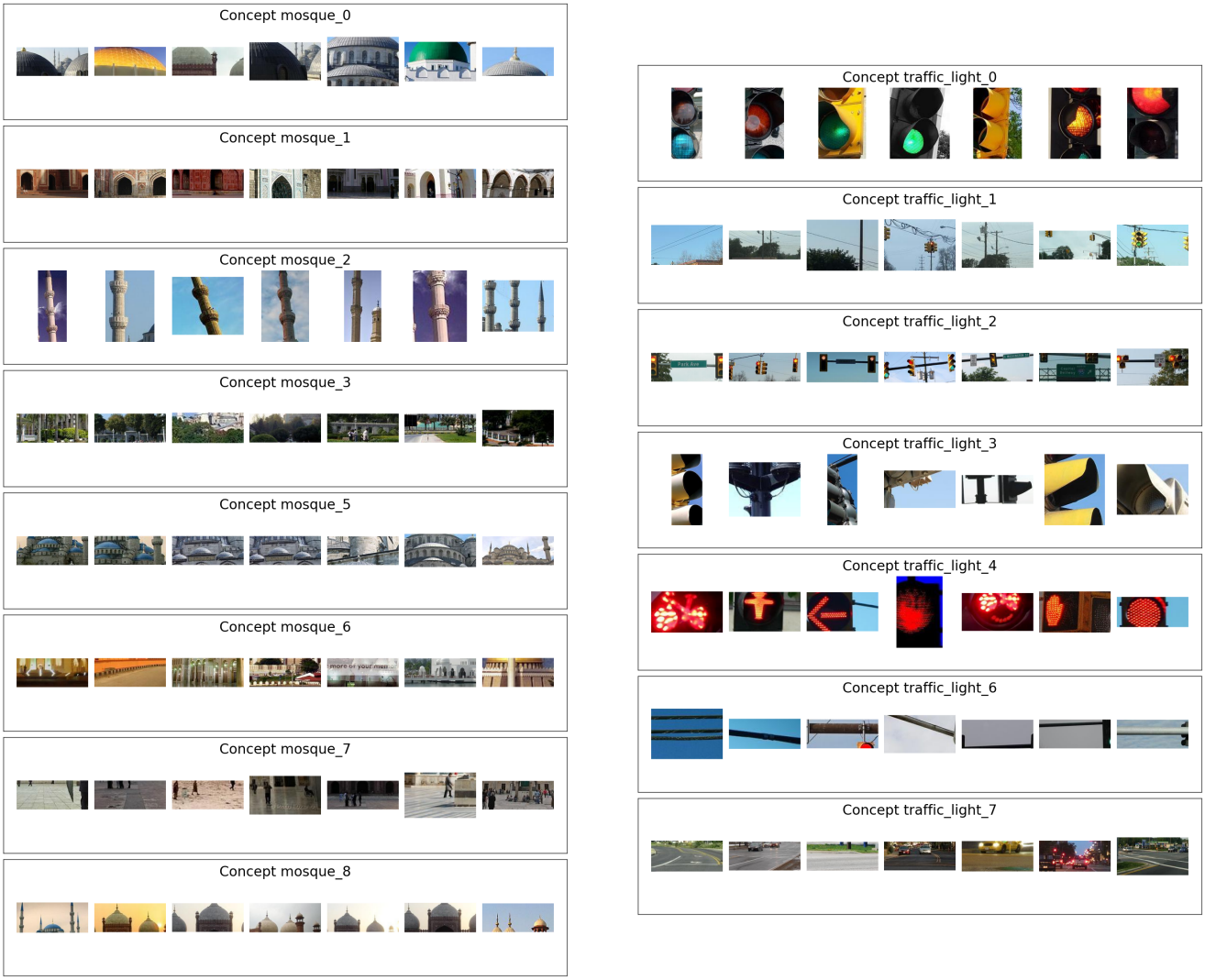


Figure 11. Concepts extracted for class “mosque” and “traffic light” with ResNet50.

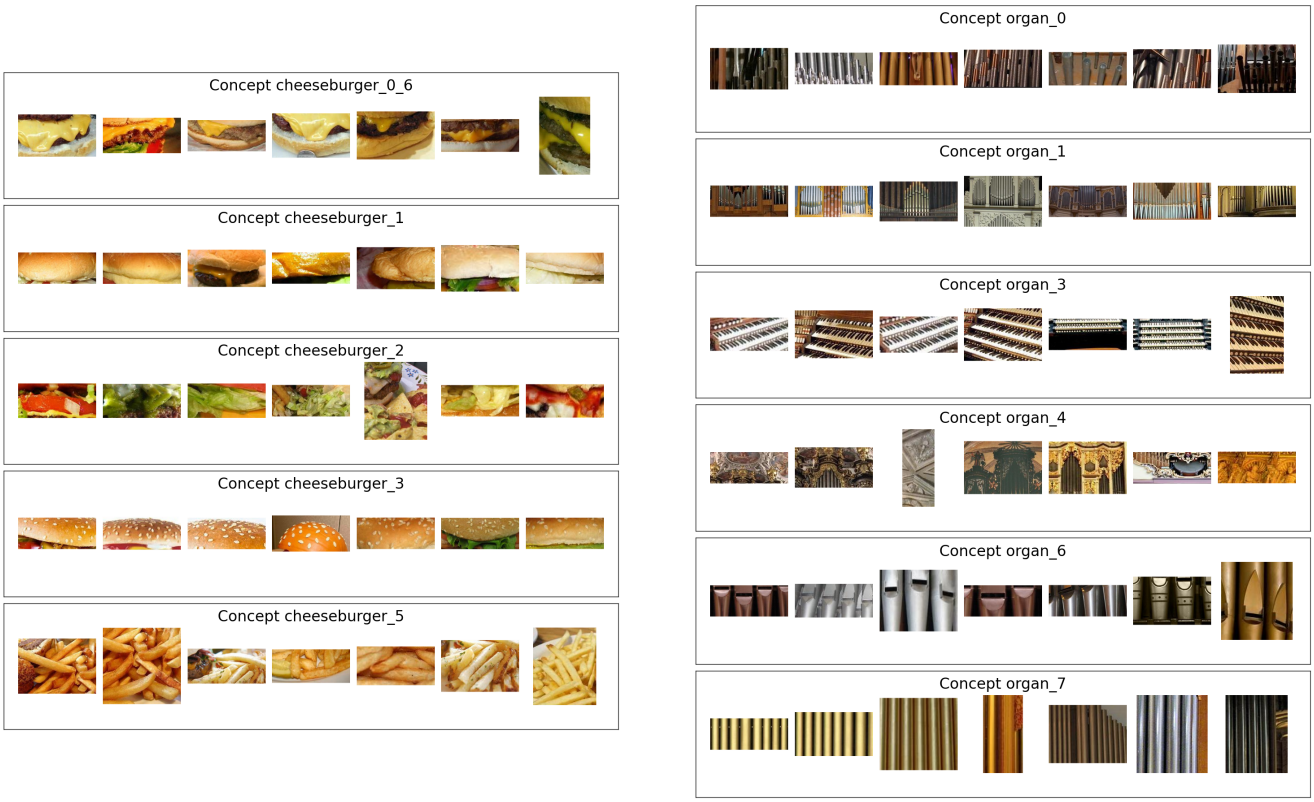


Figure 12. Concepts extracted for class “cheeseburger” and “organ” with VGG-16.

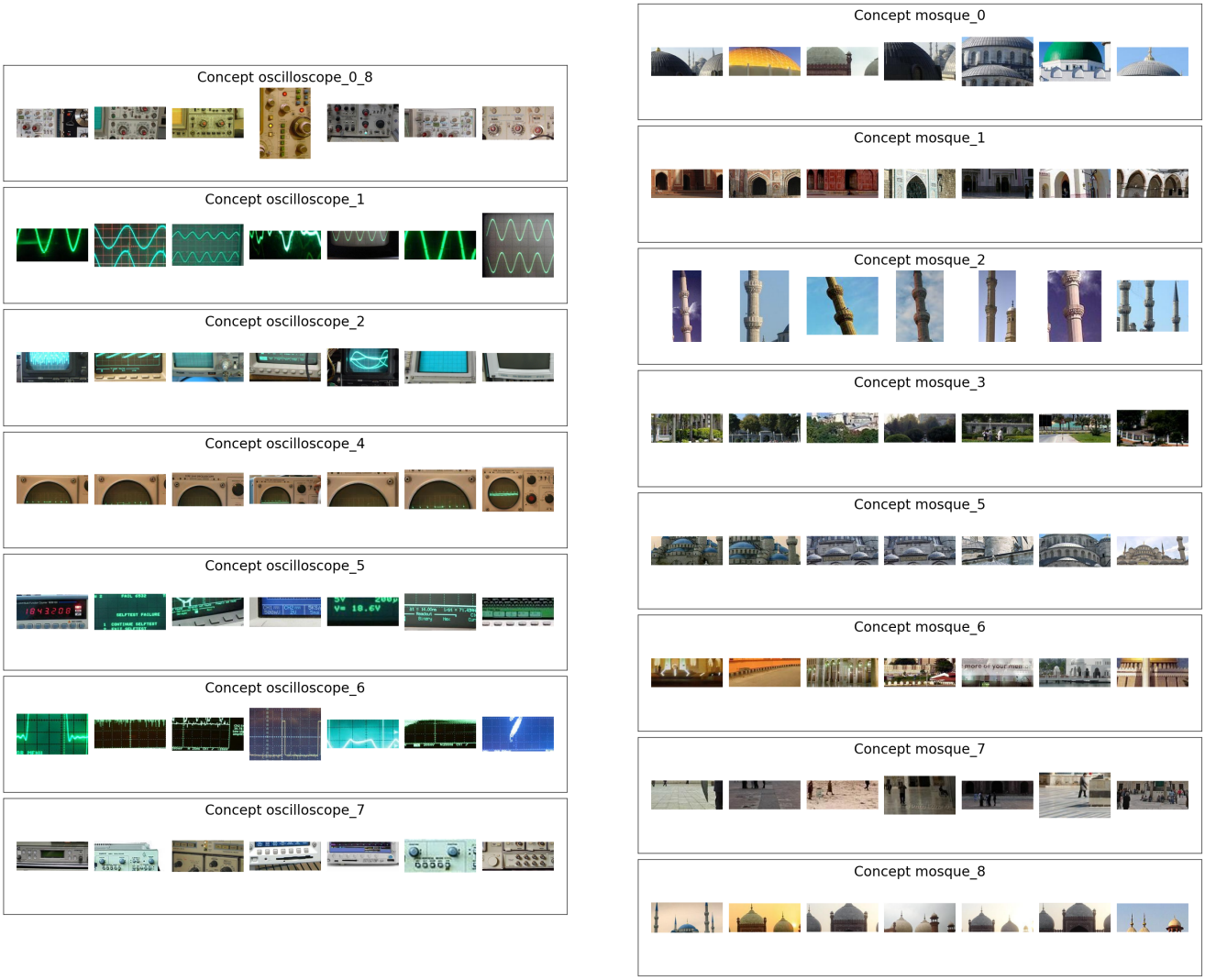


Figure 13. Concepts extracted for class “oscilloscope” and “mosque” with VGG-16.

8. Examples of global explanations

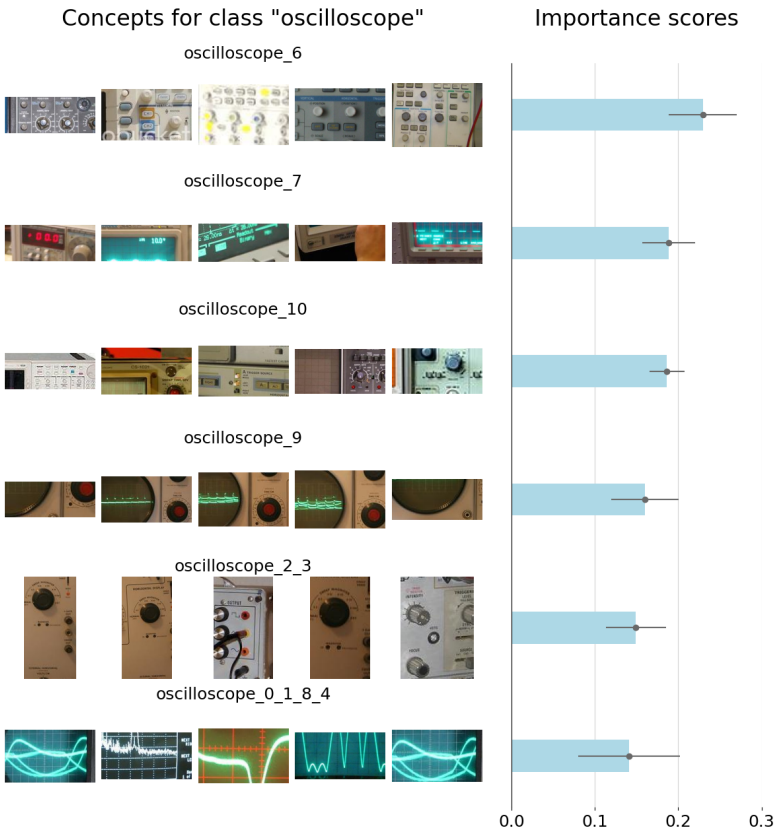


Figure 14. Global explanation for class “oscilloscope” with InceptionV3. The explanation was generated using Visual-TCAV and iterating over a set of 100 “oscilloscope” images from ImageNet.

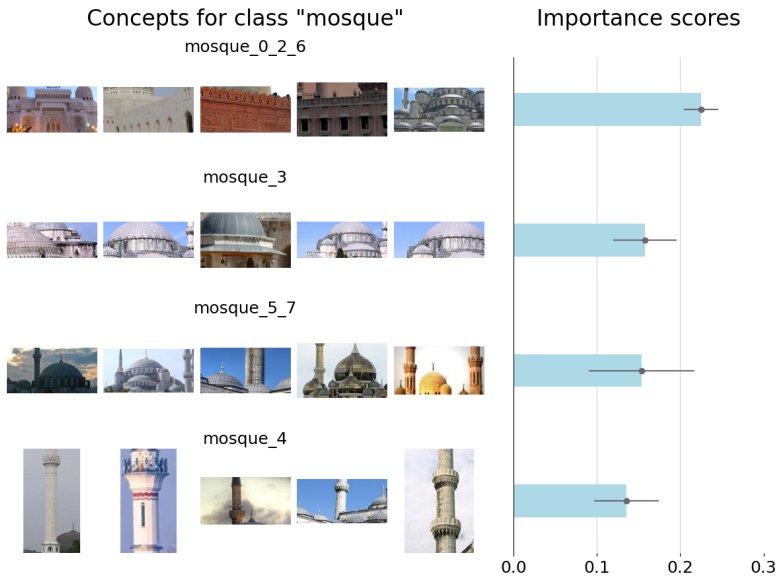


Figure 15. Global explanation for class “mosque” with InceptionV3. The explanation was generated using Visual-TCAV and iterating over a set of 100 “mosque” images from ImageNet.

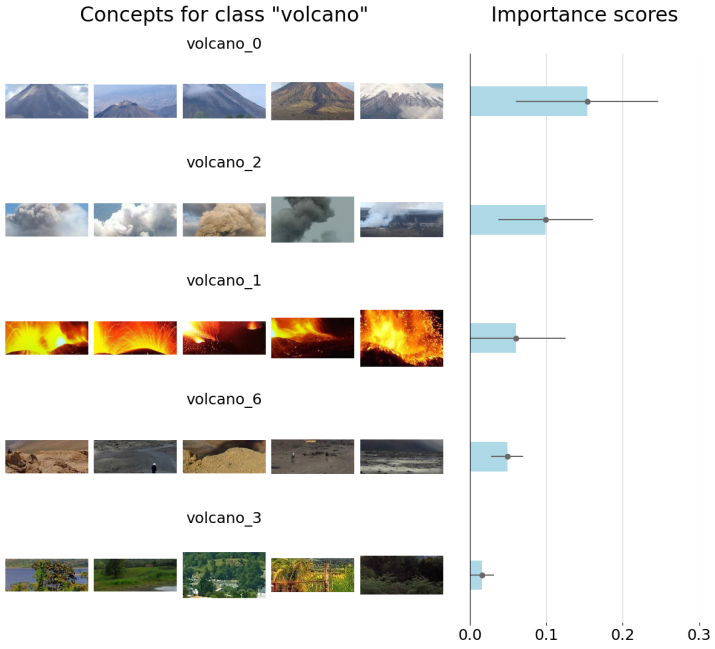


Figure 16. Global explanation for class “volcano” with ResNet50. The explanation was generated using Visual-TCAV and iterating over a set of 100 “volcano” images from ImageNet.

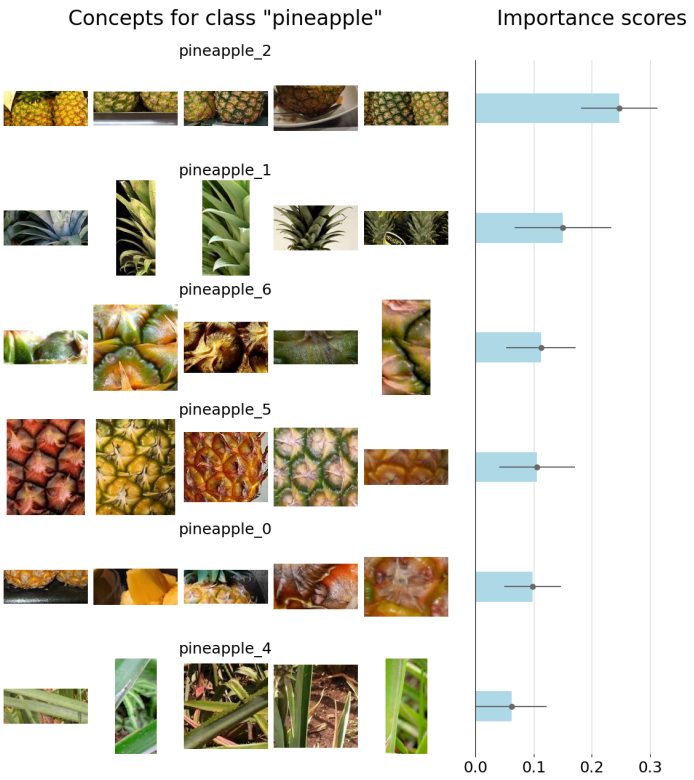


Figure 17. Global explanation for class “pineapple” with ResNet50. The explanation was generated using Visual-TCAV and iterating over a set of 100 “pineapple” images from ImageNet.

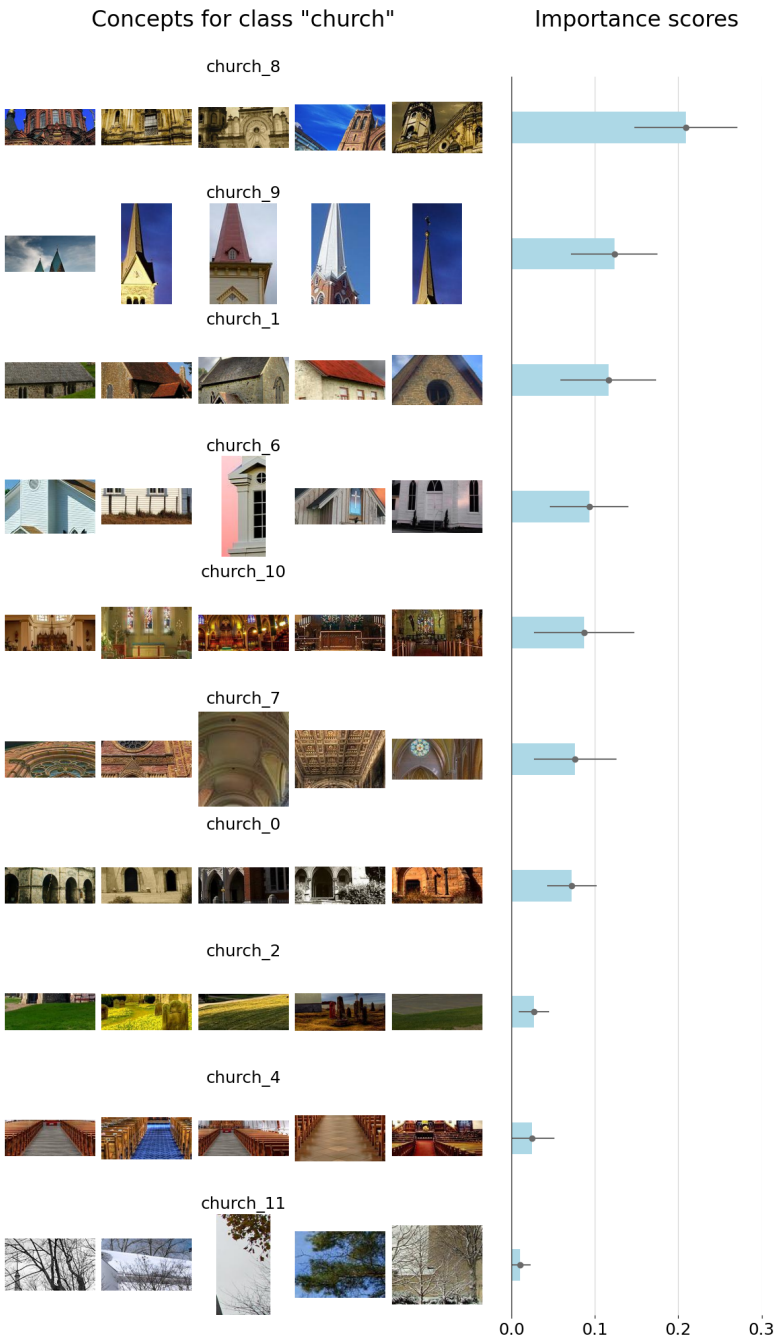


Figure 18. Global explanation for class “church” with ResNet50. The explanation was generated using Visual-TCAV and iterating over a set of 100 “church” images from ImageNet.

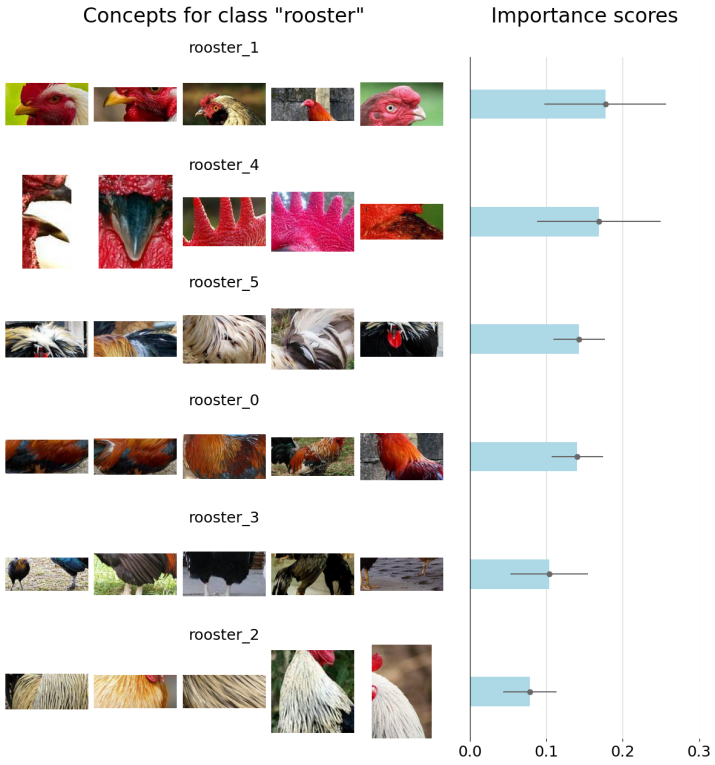


Figure 19. Global explanation for class “rooster” with VGG-16. The explanation was generated using Visual-TCAV and iterating over a set of 100 “rooster” images from ImageNet.

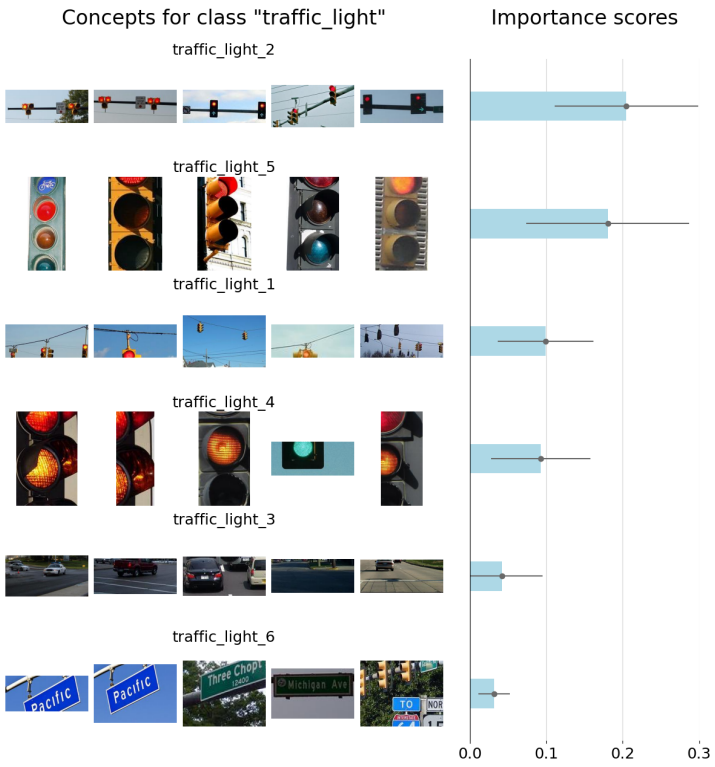


Figure 20. Global explanation for class “traffic light” with VGG-16. The explanation was generated using Visual-TCAV and iterating over a set of 100 “traffic light” images from ImageNet.

9. Examples of Local Explanations

(a) Image of class "pineapple" from ImageNet.

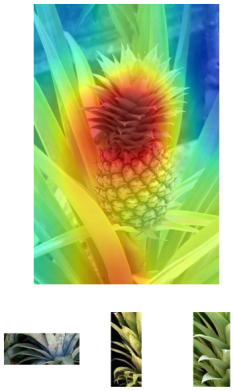
pineapple_2 concept

Class	Attrib.
pineapple	**0.27**
ear	**0.058**
banana	**0.077**

(b) Most important concept in local explanation.

pineapple_1 concept

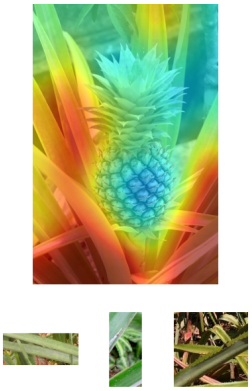
Class	Attrib.
pineapple	**0.22**
ear	**0.069**
banana	**0.045**



(c) Second most important concept in local explanation.

pineapple_4 concept

Class	Attrib.
pineapple	**0.16**
ear	**0.11**
banana	**0.052**



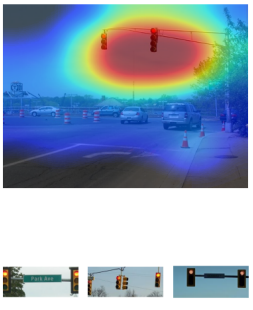
(d) Third most important concept in local explanation.

Figure 21. Local explanation of an image of class "pineapple", generated with Visual-TCAV using the concepts extracted for class "pineapple" with ResNet50.

(a) Image of class "traffic light" from ImageNet.

traffic_light_2 concept

Class	Attrib.
traffic_light	**0.18**
street_sign	**0.076**
trolleybus	**0.04**



(b) Most important concept in local explanation.

traffic_light_7 concept

Class	Attrib.
traffic_light	**0.14**
street_sign	**0.12**
trolleybus	**0.14**

(c) Second most important concept in local explanation.

traffic_light_1 concept

Class	Attrib.
traffic_light	**0.11**
street_sign	**0.068**
trolleybus	**0.071**

(d) Third most important concept in local explanation.

Figure 22. Local explanation of an image of class "traffic light", generated with Visual-TCAV using the concepts extracted for class "traffic light" with ResNet50.

church_5 concept

Class	Attrib.
church	**0.26**
bell_cote	**0.18**
monastery	**0.24**

church_1 concept

Class	Attrib.
church	**0.23**
bell_cote	**0.11**
monastery	**0.13**

church_6 concept

Class	Attrib.
church	**0.19**
bell_cote	**0.13**
monastery	**0.14**

(a) Image of class “church” from ImageNet.

(b) Most important concept in local explanation.

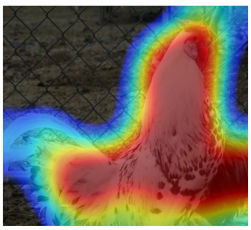
(c) Second most important concept in local explanation.

(d) Third most important concept in local explanation.

Figure 23. Local explanation of an image of class “church”, generated with Visual-TCAV using the concepts extracted for class “church” with VGG-16.

rooster_5 concept

Class	Attrib.
rooster	**0.16**
hen	**0.13**
partridge	**0.092**

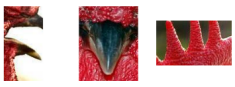


rooster_1 concept

Class	Attrib.
rooster	**0.14**
hen	**0.12**
partridge	**0.067**

rooster_4 concept

Class	Attrib.
rooster	**0.13**
hen	**0.12**
partridge	**0.058**



(a) Image of class “rooster” from ImageNet.

(b) Most important concept in local explanation.

(c) Second most important concept in local explanation.

(d) Third most important concept in local explanation.

Figure 24. Local explanation of an image of class “rooster”, generated with Visual-TCAV using the concepts extracted for class “rooster” with VGG-16.

10. Examples Questions from the Concept Matching Game

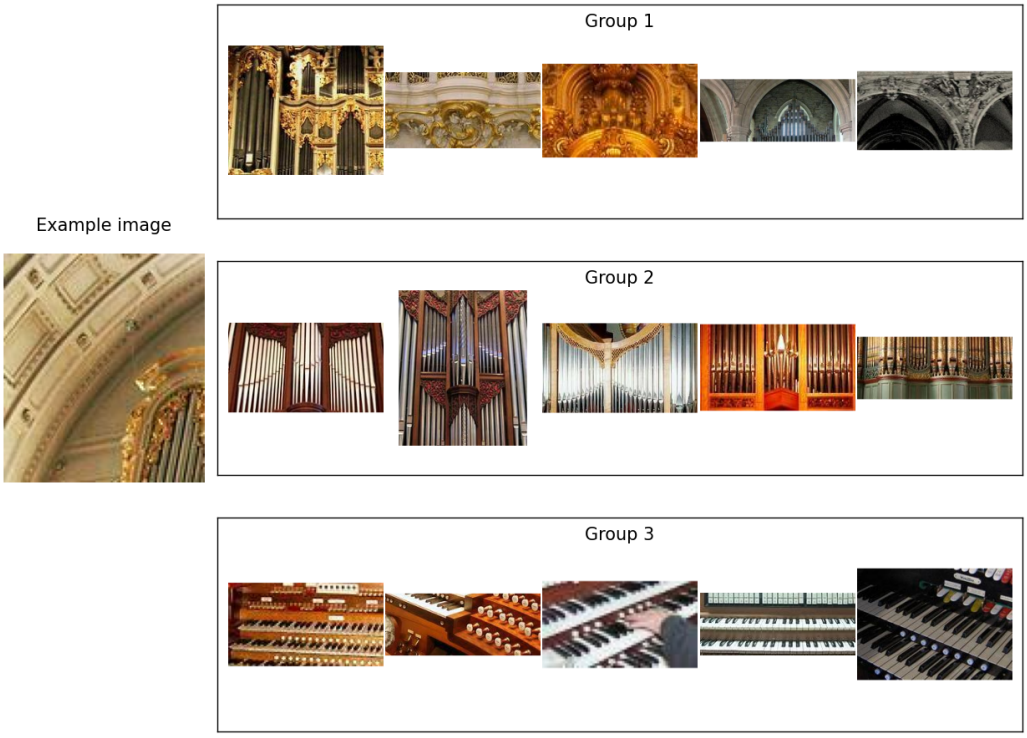


Figure 25. A sample question from the concept matching game, presenting an option for the class “organ”.

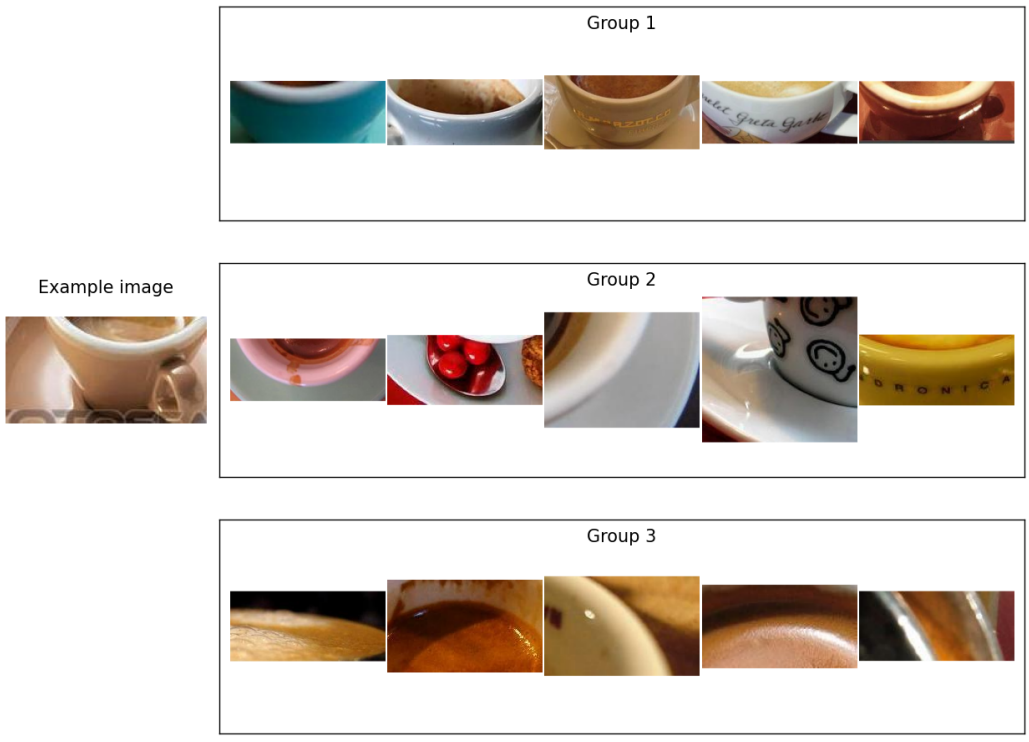
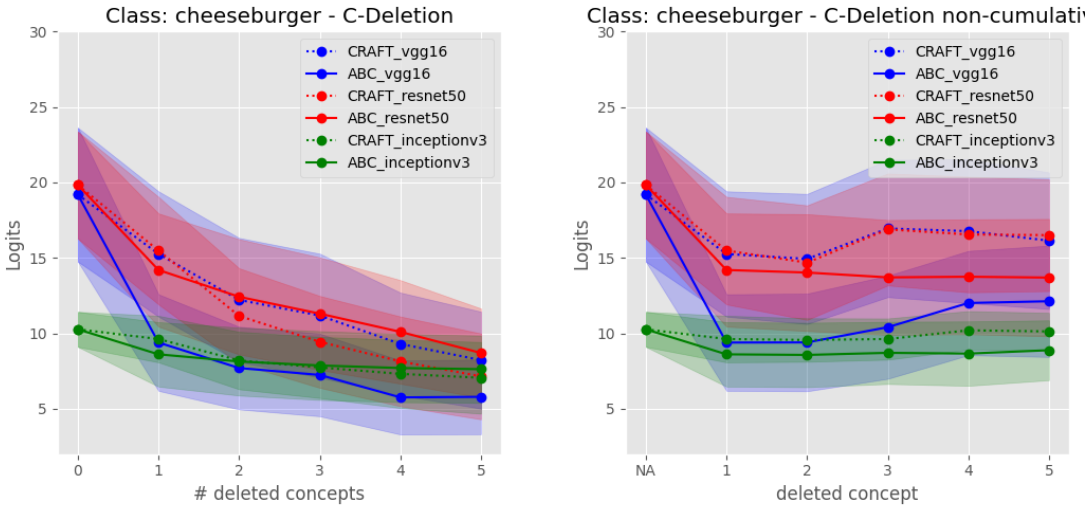


Figure 26. An example question from the concept matching game, showing a choice for class “espresso”.

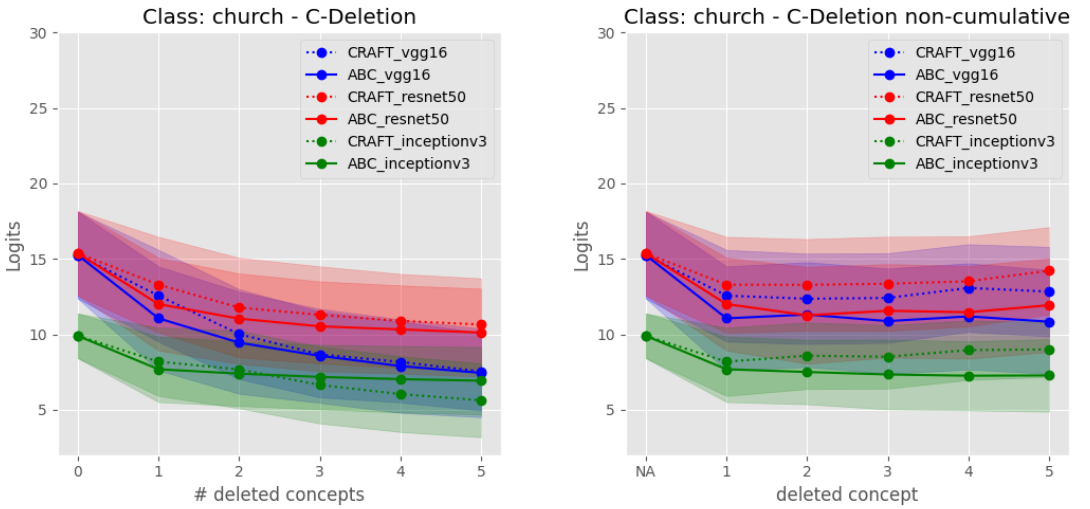
11. Fidelity Study's C-Deletion Plots



(a) Cumulative c-deletion plot.

(b) Non-cumulative c-deletion plot.

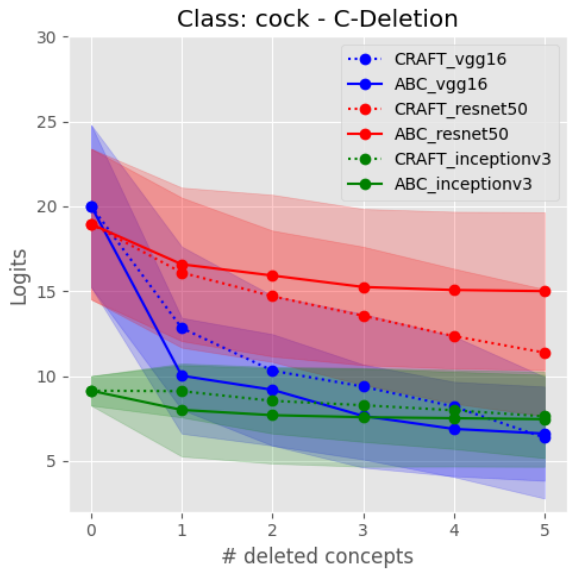
Figure 27. C-deletion results comparing ABC and CRAFT for class "cheeseburger" using 500 images.



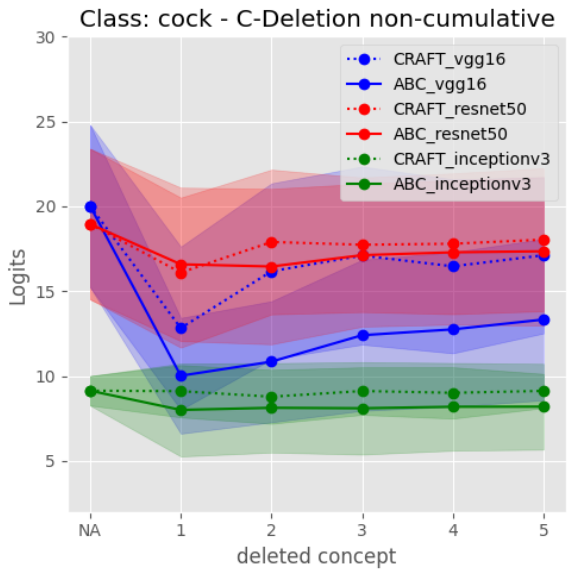
(a) Cumulative c-deletion plot.

(b) Non-cumulative c-deletion plot.

Figure 28. C-deletion results comparing ABC and CRAFT for class "church" using 500 images.

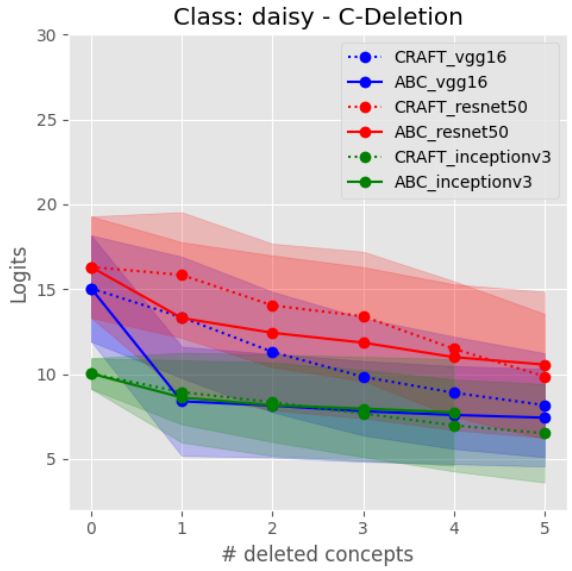


(a) Cumulative c-deletion plot.

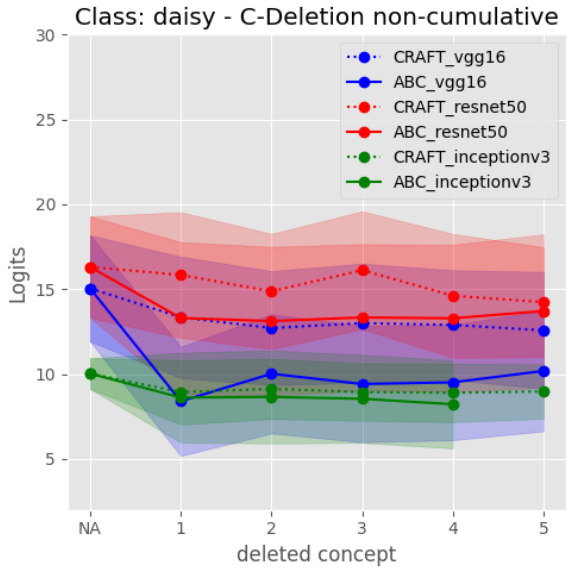


(b) Non-cumulative c-deletion plot.

Figure 29. C-deletion results comparing ABC and CRAFT for class “cock” using 500 images.

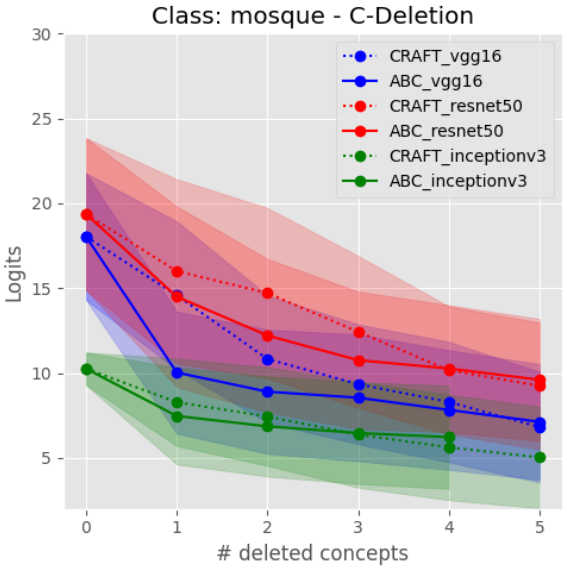


(a) Cumulative c-deletion plot.

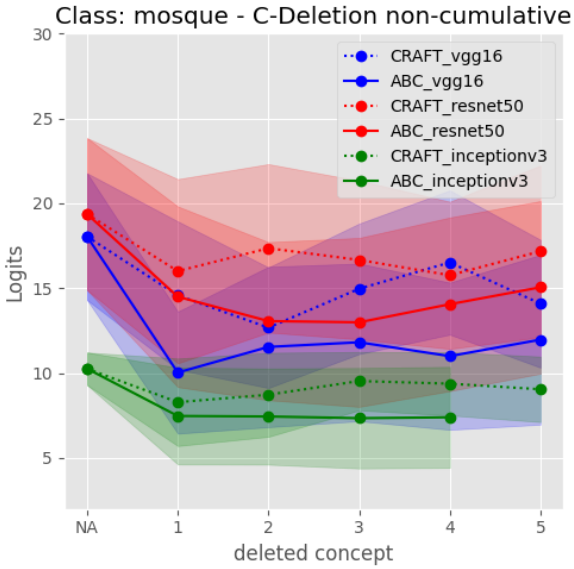


(b) Non-cumulative c-deletion plot.

Figure 30. C-deletion results comparing ABC and CRAFT for class “daisy” using 500 images.

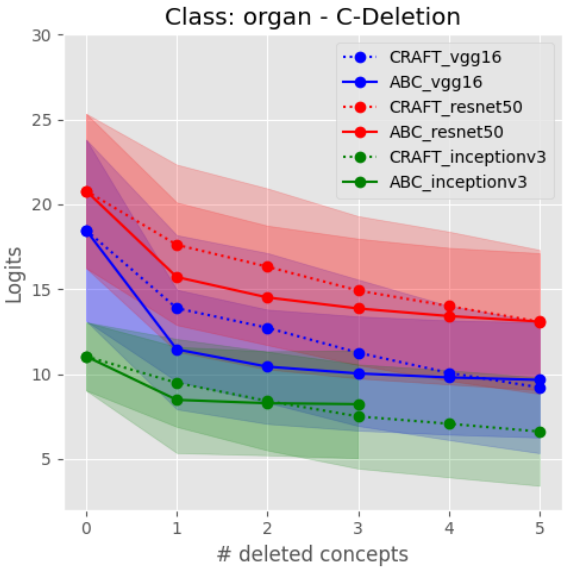


(a) Cumulative c-deletion plot.

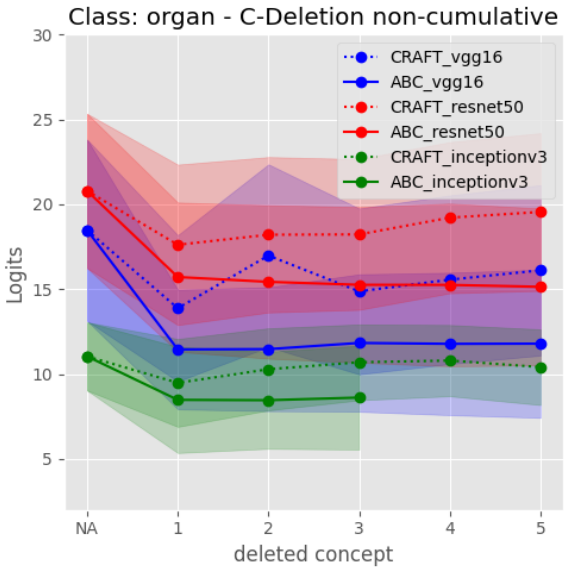


(b) Non-cumulative c-deletion plot.

Figure 31. C-deletion results comparing ABC and CRAFT for class “mosque” using 500 images.

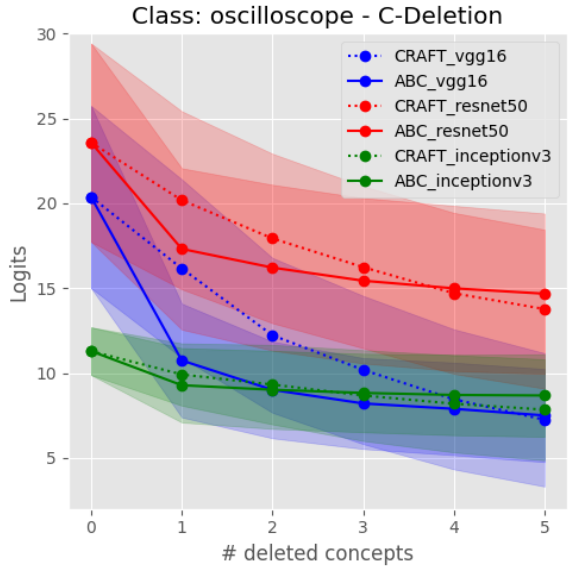


(a) Cumulative c-deletion plot.

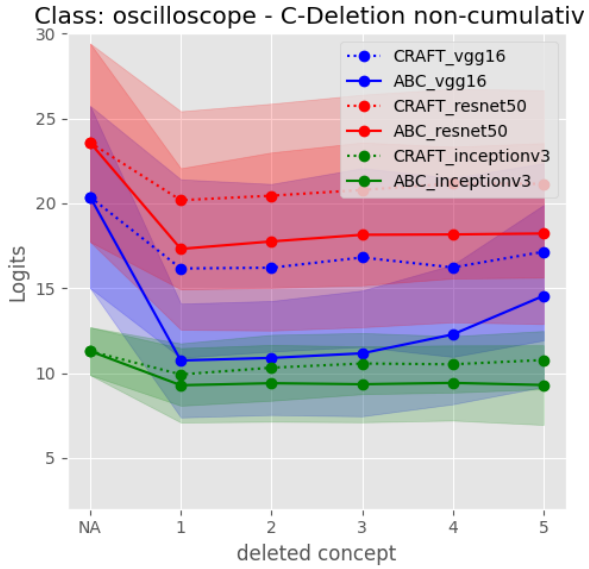


(b) Non-cumulative c-deletion plot.

Figure 32. C-deletion results comparing ABC and CRAFT for class “organ” using 500 images.

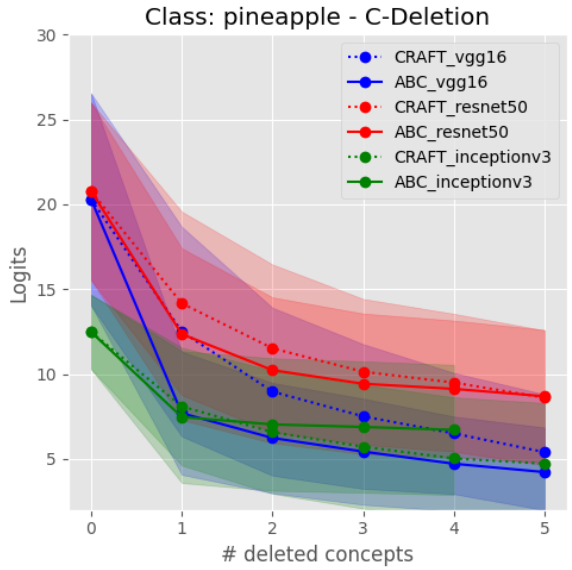


(a) Cumulative c-deletion plot.

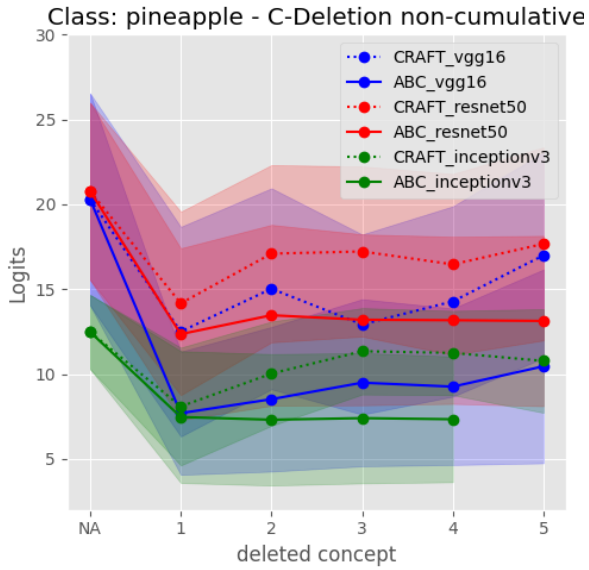


(b) Non-cumulative c-deletion plot.

Figure 33. C-deletion results comparing ABC and CRAFT for class “oscilloscope” using 500 images.

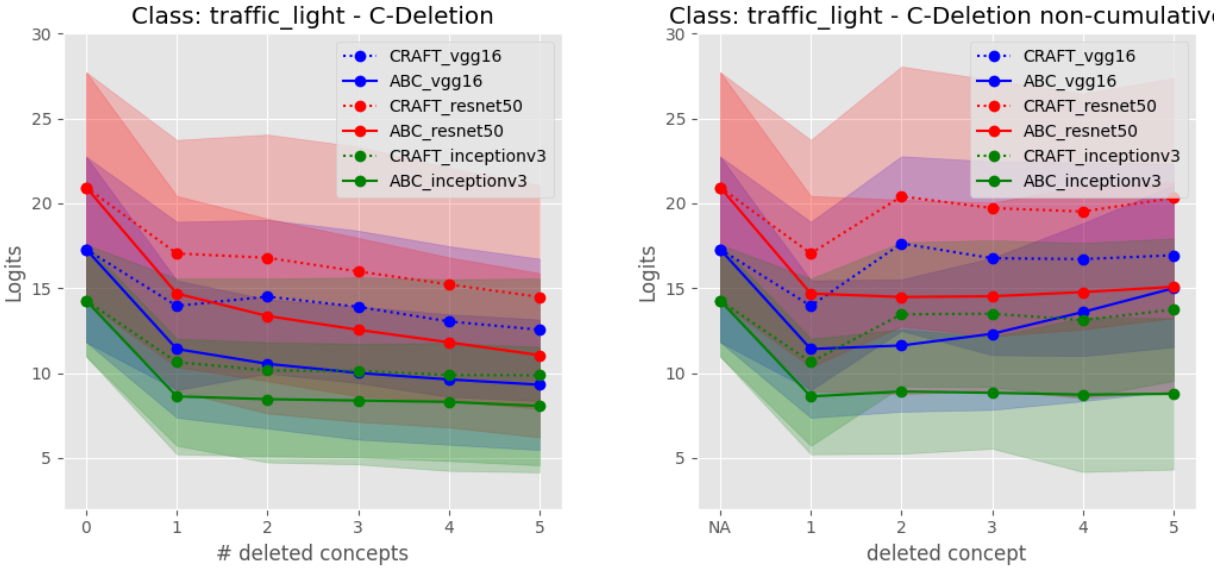


(a) Cumulative c-deletion plot.



(b) Non-cumulative c-deletion plot.

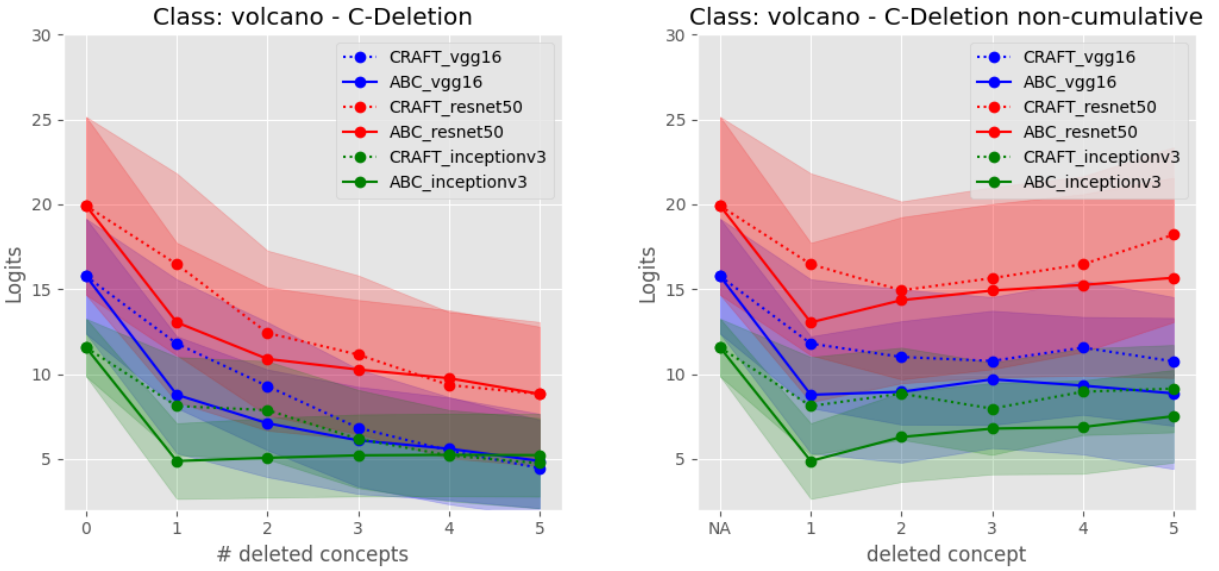
Figure 34. C-deletion results comparing ABC and CRAFT for class “pineapple” using 500 images.



(a) Cumulative c-deletion plot.

(b) Non-cumulative c-deletion plot.

Figure 35. C-deletion results comparing ABC and CRAFT for class "traffic_light" using 500 images.



(a) Cumulative c-deletion plot.

(b) Non-cumulative c-deletion plot.

Figure 36. C-deletion results comparing ABC and CRAFT for class "volcano" using 500 images.